Technical Report # 39

The Development of Early Literacy Measures for use in a Progress Monitoring Assessment System: Letter Names, Letter Sounds and Phoneme Segmenting

Julie Alonzo

Gerald Tindal

University of Oregon

Published by

Behavioral Research and Teaching
University of Oregon • 175 Education
5262 University of Oregon • Eugene, OR 97403-5262
Phone: 541-346-3535 • Fax: 541-346-5689
http://brt.uoregon.edu

Introduction

In this technical report, we describe the development alternate forms of three types of early literacy measures as part of a comprehensive progress monitoring literacy assessment system developed in 2006 for use with students in Kindergarten through fourth grade. We begin with a brief overview of the two conceptual frameworks underlying the assessment system: progress monitoring and developmental theories of reading. We then provide context for how the three early literacy measures of Letter Names, Letter Sounds, and Phoneme Segmenting measures fit into the full assessment system. Subsequent technical reports provide similar information about measures of Word and Passage Reading Fluency (Alonzo & Tindal, 2007), Passage, and Reading Comprehension (Alonzo, Liu, & Tindal, 2007).

Conceptual Framework: Progress Monitoring and Literacy Assessment

Early work related to curriculum-based measurement (CBM) led by Deno and Mirkin at the University of Minnesota (*c.f.a.*, Deno & Mirkin, 1977) was instrumental in promoting the use of short, easily-administered assessments to provide educators with information about student skill development useful for instructional planning. In the three decades since, such *progress monitoring probes* as they have come to be called have increased in popularity, and they are now a regular part of many schools' educational programs (Alonzo, Ketterlin-Geller, & Tindal, 2007). However, CBMs – even those widely used across the United States – often lack the psychometric properties expected of modern technically-adequate assessments. Although the precision of instrument development has advanced tremendously in the past 30 years with the advent of more sophisticated statistical techniques for analyzing tests on an item by item basis rather than relying exclusively on comparisons of means and standard deviations to evaluate comparability of alternate forms, the world of CBMs has not always kept pace with these statistical advances.

A key feature of assessments designed for progress monitoring is that alternate forms must be as equivalent as possible to allow meaningful interpretation of student performance data across time. Without such cross-form equivalence, changes in scores from one testing session to the next are difficult to attribute to changes in student skill or knowledge. Improvements in student scores may, in fact, be an artifact of the second form of the assessment being easier than the form that was administered first. The advent of more sophisticated data analysis techniques (such as the Rasch modeling used in this study) have made it possible to increase the precision with which we develop and evaluate the quality of assessment tools. In this technical report, we document the development of a progress monitoring assessment in reading, designed for use with students in Kindergarten through Grade 4. This assessment system was developed to be used by elementary school educators interested in monitoring the progress their students make in the area of early reading skill acquisition.

Reading is a somewhat fluid construct, shifting over time from a focus on discrete skills necessary for working with language in both written and spoken forms, to those more complex combinations of skills associated with decoding, and finally to comprehension—a construct in which all prior literacy skills are called upon in the act of reading. Reading assessment typically follows this general progression as well (Reading First, 2006). Assessments of emerging literacy skills evaluate student mastery of the alphabetic principal. These tests measure students' ability to correctly identify and/or produce letters and the sounds associated with them. They measure students' ability to manipulate individual phonemes (sound units) within words, when, for example, students are asked to blend a list of phonemes into a word, segment a word into its corresponding phonemes, or identify the sounds which begin or end a word (Ritchey & Speece, 2006). The relationships between these constructs in English are well-documented in the

research literature. In early readers, ability to identify letter names and the sounds that letters make predicts phonemic awareness. Phonemic awareness predicts fluency, and low fluency is a strong predictor of difficulties in reading (National Reading Panel, 2000).

As student reading skill progresses, it is necessary to use different reading measures to be able to continue to track the progress students are making as developing readers. Oral reading fluency, which measures a combination of students' sight vocabulary and their ability to decode novel words rapidly and accurately, is consistently identified in the literature as one of the best predictors of student reading comprehension in the early grades (Graves, Plasencia-Peinado, Deno, & Johnson, 2005; Hasbrouck & Tindal, 2005). Eventually, however, the information provided by measures of oral reading fluency is limited. Readers attain a fluency threshold that enables them to attend to comprehension rather than decoding (Ehri, 1991, 2005). Once this threshold has been reached, fluency is no longer sensitive to increases in reading comprehension. At this point, one must turn to measures designed to assess comprehension more directly. Although this technical report provides information specifically related to the Letter Names measures developed for use in our Progress Monitoring assessment system, it is important to provide an overview of the complete system so readers can understand how the Letter Names measures fit into the system as a whole.

The Measures that Comprise Our Complete Assessment System

Based on previous empirical studies of early literacy assessment (see, for example, the report from the National Reading Panel, 2000), we decided to develop two measures of alphabetic principle (Letter Names and Letter Sounds), one measure of Phonological Awareness (Phoneme Segmenting), two measures of fluency (Word Reading Fluency and Passage Reading Fluency), and one measure of comprehension (Multiple Choice Reading Comprehension). The

specific technical specifications for the Letter Names, Letter Sounds, and Phoneme Segmenting measures are described in the methods section of this technical report. First, we describe the specific requirements related to the intended use of the measures in our assessment system.

When one is interested in monitoring the progress students are making in attaining specific skills, it is important to have sufficient measures to sample student performance frequently. Thus, our goal was to create 20 alternate forms of each measure in our assessment system at each grade level where the measure was designed to be used (see Table 1). Because these alternate forms are designed to be used for progress monitoring, it is essential that all forms of a particular measure in a given grade level be both sensitive to showing growth in a discrete skill area over short periods of time (1-2 weeks of instruction) and comparable in difficulty. These two equally important needs informed all parts of our measurement development effort: the construction of the technical specifications for each of the measures, the design of the studies used to gather data on item and test functioning, the analytic approaches we used to interpret the results of the pilot studies, and subsequent revision of the measures. In all cases, we sought approaches that would provide us with enough information to evaluate the *sensitivity of the individual measures* to detect small differences in student performance and the *comparability of the different forms* of each measure to allow for meaningful interpretation of growth over time.

Table 1

Distribution of the Measures Across the Grades

Grade	Measure					
	Letter Names	Letter Sounds	Phoneme Segmenting	Word Reading Fluency	Passage Reading Fluency	MC Reading Comp
Kindergarten	X*	X	X	X		
Grade 1	X	X	X	X	X	
Grade 2			X	X	X	
Grade 3				X	X	X
Grade 4					X	X

*Note: Each "X" represents 20 alternate forms of the measure for that grade level.

In the section that follows, we describe the piloting methods used to gather information on the relative difficulty of the different letter names, letter sounds, and words used in phoneme segmenting measures allowing us to create an item bank from which we could draw to construct 20 comparable alternate forms of these types of early literacy measures for use in a progress monitoring assessment system.

The Letter Names Measure

The Letter Names measure tests students' ability to name the letters of the English alphabet, both in their lower case and capitalized forms. In this individually-administered measure, students are shown a series of letters organized in a chart on one side of a single sheet of paper and given a set amount of time (ranging from 30 – 60 seconds on different versions of this measure) to name as many of them as they can. A trained assessor follows along as the student names the letters, indicating on his/her own test protocol each letter the student reads incorrectly and prompting the student to go on if he/she hesitates at a letter for more than three

seconds. Student self-corrections are counted as correct responses. At the end of the allotted time, the assessor marks the last letter named and calculates the total number of letters read correctly to arrive at the student's score, letters named correctly in one minute (on tests administered for shorter times, it is common practice to use a multiplier to convert the raw score to a 'per minute' fluency-based score).

The Letter Sounds Measure

The Letter Sounds measure tests students' ability to identify the sounds associated with the letters of the English alphabet, both in their lower case and capitalized forms. In this individually-administered measure, students are shown a series of letters organized in a chart on one side of a single sheet of paper and given a set amount of time (ranging from 30 – 60 seconds on different versions of this measure) to name as many of them as they can. A trained assessor follows along as the student produces the sounds associates with each of the letters, indicating on his/her own test protocol each letter for which the student fails to correctly identify the sound and prompting the student to go on if he/she hesitates at a letter for more than three seconds. Student self-corrections are counted as correct responses. At the end of the allotted time, the assessor marks the last letter sound produced and calculates the total number of letters sounds produced correctly to arrive at the student's score, letter sounds produced correctly in one minute (on tests administered for shorter times, it is common practice to use a multiplier to convert the raw score to a 'per minute' fluency-based score).

The Phoneme Segmenting Measure

The Phoneme Segmenting measure tests students' ability to segment a word into its constituent phonemes. In this individually-administered measure, test administrators follow a standard written protocol on which is listed a series of words. They say each word aloud, asking

students to segment the word into its individual sounds. As students finish segmenting one word, test administrators provide the next word verbally, repeating this sequence for a set amount of time (typically 60 seconds) to segment as many words into phonemes as they can. As students say the phonemes, assessors indicate on their own test protocol each phoneme the student correctly segments. Student self-corrections are counted as correct responses, and students are prompted to go on if they hesitate for more than three seconds. At the end of the allotted time, the assessor marks the last phoneme produced and calculates the total number of phonemes segmented correctly to arrive at the student's score, phonemes segmented correctly in one minute (on tests administered for shorter times, it is common practice to use a multiplier to convert the raw score to a 'per minute' fluency-based score).

Methods

Our goal was to create 20 alternate forms of each measure at each grade level where the measure was designed to be used (see Table 1). Because these alternate forms will be used for progress monitoring, it is essential that all forms of a particular measure in a given grade level be comparable in difficulty. To design alternate forms of the measures, we gathered information about the difficulty of specific test items during a pilot testing session between May 15 and June 9, 2006 and created an item pool from which we could draw as we created the 20 alternate forms of the three types of measures.

Research Design

Following recommendations by Kolen and Brennan (1995), we used a common item nonequivalent groups design to pilot each item on all three types of measures. We used information from this piloting to create a calibrated item pool, with all letter names placed on the same Θ scale. In other words, although not every letter name was administered to every student

in the pilot testing, there was enough overlap of items between the different forms of the measure used in piloting to allow us to analyze all the data simultaneously. The items shared across the different forms of the test (common items) allowed us to calibrate all items on the same metric, an essential pre-requisite for creating an item bank.

Piloting the Letter Names items. To accomplish our ultimate goal of being able to create 20 comparable alternate forms of the Letter Names measure in Kindergarten and first grade, we gathered information about the difficulty of each letter in its capital and lower case format by administering both forms of each letter to a sample of kindergarten and first-grade students in a large suburban school district in the Pacific Northwest. In all, between 297 and 1036 students provided pilot test data on each letter, with the letters used as anchor items accounting for the highest number of student interactions.

To reduce the likelihood that fatigue would influence student performance on the measure of letter names, resulting in less reliable item information, we gathered item-level data on only the first two lines (20 letters) of each full-length Letter Names test. We created three different forms of the Letter Names test, randomly seeding all letters in their capital and lower case formats across all three forms of the test, retaining 5 letters as anchor items, common across all three forms of the test. No exact letters were repeated in these two rows until all other letter possibilities had been used. For the purposes of our test specifications “exact letters” were defined as a letter in either its capital or lower case format. Thus, a lower case ‘b’ might appear on a test form before all other letters of the alphabet had been used even if a capital “B” had already appeared on the test form. To allow for later equating and scaling across and between forms, the five anchor item letters appeared consistently in the same locations on all forms of the Letter Names measure (see Figure 1).

Figure 1
Letter Names Template, Showing Locations of Five Items Common to All Forms

		s			N	p		h	
g									

These anchor items were used during analysis to allow concurrent estimation of item difficulty across all three forms of the test. In keeping with Kolen and Brennan's (1995) recommendations, roughly 20% of the items overlapped from one form to another, and the anchor items were located in the same position on each form of the test.

Piloting the Letter Sounds items. To accomplish our ultimate goal of being able to create 20 comparable alternate forms of the Letter Sounds measure in Kindergarten and first grade, we gathered information about the difficulty of each letter sound in its capital and lower case format by administering both forms of each letter sound to a sample of kindergarten and first-grade students in a large suburban school district in the Pacific Northwest. In all, between 554 and 1801 students provided pilot test data on each letter sound, with the letter sounds used as anchor items accounting for the highest number of student interactions.

To reduce the likelihood that fatigue would influence student performance on the measure of letter names, resulting in less reliable item information, we gathered item-level data on only the first two lines (20 items) of each full-length Letter Sounds test. We created three different forms of the Letter Sounds test, randomly seeding all letters in their capital and lower case formats across all three forms of the test, retaining 5 letters as anchor items, common across all three forms of the test. No exact letters were repeated in these two rows until all other letter possibilities had been used. For the purposes of our test specifications "exact letters" were defined as a letter or digraph (combination of two consonants that form a single phoneme) in either its capital or lower case format. Thus, a lower case 'b' might appear on a test form before all other letters of the alphabet had been used even if a capital "B" had already appeared on the test form. To allow for later equating and scaling across and between forms, the five anchor item letters appeared consistently in the same locations on all forms of the Letter Sounds measure (see Figure 2).

Figure 2
Letter Sounds Template, Showing Locations of Five Items Common to All Forms

		b			h	r		K	
qu									

These anchor items were used during analysis to allow concurrent estimation of item difficulty across all three forms of the test. In keeping with Kolen and Brennan's (1995) recommendations, roughly 20% of the items overlapped from one form to another, and the anchor items were located in the same position on each form of the test.

Piloting the Phoneme Segmenting items. To accomplish our ultimate goal of being able to create 20 comparable alternate forms of the Phoneme Segmenting measure in Kindergarten and first grade, we gathered information about the difficulty of each word on the measure by administering different forms of the test to a sample of kindergarten and first-grade students in a large suburban school district in the Pacific Northwest. In all, between 110 and 2067 students provided pilot test data on words used in the Phoneme Segmenting measures, with the words used as anchor items accounting for the highest number of student interactions.

To allow for later equating and scaling across and between forms, the five anchor item words appeared consistently in the same locations on all forms of the Phoneme Segmenting measure (see Figure 3).

Figure 3
Phoneme Segmenting Template, Showing Locations of Five Items Common to All Forms

#	Item
1	
2	tap
3	
4	nine
5	
6	cup

7	
8	city
9	
10	show

These anchor items were used during analysis to allow concurrent estimation of item difficulty across all forms of the test. In keeping with Kolen and Brennan's (1995) recommendations, roughly 20% of the items overlapped from one form to another, and the anchor items were located in the same position on each form of the test.

Analysis

Item parameters were estimated using a one-parameter Rasch model analyzed with Winsteps3.61.1 analytic software (Linacre, 2006). Rasch analyses differ from approaches using classical statistics in that they consider patterns of responses across individuals, using this information to provide a level of specificity in results unattainable with approaches based on classical statistics used in the development of most CBMs. In a complex iterative process, a Rasch analysis concurrently estimates the difficulty of individual test items and the ability level of each individual test taker. The results one obtains from this analysis, relevant to our discussion here, include an estimation of the difficulty (referred to as the *measure* of each item), the *standard error of measure* associated with each item's estimated difficulty, and the degree to which each item 'fits' the measurement model (referred to as the *mean square outfit* of each item). All of this information must be considered when evaluating the technical adequacy of the measures, as described below.

Considering each item's estimated difficulty. Rasch analyses, which examine each item's reliability, provide a more precise treatment of reliability than classical statistics, which examine the issue only at a more global test level. The most reliable estimation of a test-taker's ability can

be gained from tests comprised of items that represent the fullest range of difficulty possible for the population with which the test is intended to be used. Thus, in creating our Letter Names measures, it is necessary for us to select items representing a range of difficulties. In Rasch analyses, this information is gleaned from examining each item's *measure*. Easy items will have measures represented with negative numbers; difficult items will have measures represented with positive numbers. A measure of zero indicates an item that a person of average ability would be expected to have a 50% chance of getting correct. Thus, we sought a full range of measures on every Letter Names measure.

Examining the standard error of measure. Rasch analyses provide information about the standard error of measure associated with the estimation of each item's measure. In general, the smaller the standard error of measure, the more reliable the estimation. We sought small standard errors of measure on all items on our tests. Items where the standard error of measure is too great for reliable estimation are indicated on the output files with a notation that the computer program was unable to provide a reliable estimate of the item's difficulty.

Using the mean square outfit to evaluate goodness of fit. An additional piece of information used to evaluate technical adequacy in a Rasch model is the mean square outfit associated with each item. Values in the range of 0.50 to 1.50 are considered *acceptable fit*. Mean square outfits falling outside this acceptable range indicate the need for further evaluation of item functioning. Such further evaluation takes into consideration additional sources of information, such as the standard error associated with the item's estimation as well as the sample size used to generate the estimate of model fit. In general, items with a mean square outfit less than 0.50 are considered less worrisome than items with mean square outfits higher

than 1.50. Our technical specifications called for the exclusion of any items with unacceptable mean square outfits from the item bank.

Results

Because each of the three types of measures were piloted in separate studies, results will be presented for each individually.

Letter Names

The Letter Names items were piloted in the spring of 2006 using 5 common items across 3 separate forms of the measure to equate items across forms. Table 2 presents the results of this pilot testing.

Table 2

Results of IRT Analysis of Letter Names Test, Spring Pilot, 2006

Letter	Upper or Lower Case	Count	Measure	Mean Square Outfit
o	Lower case	306	-2.55	.77
X	Upper case	297	-2.14	1.65
A	Upper case	297	-1.67	.24
s	Lower case	1036	-1.64	2.36
O	Upper case	433	-1.64	.39
B	Upper case	306	-1.42	1.29
E	Upper case	433	-1.18	1.70
a	Lower case	433	-1.18	.46
T	Upper case	433	-1.18	.33
x	Lower case	297	-1.17	.55
e	Lower case	297	-.99	.86
r	Lower case	306	-.96	.73
Z	Upper case	306	-.90	1.40
S	Upper case	306	-.77	.81
L	Upper case	297	-.72	.62
t	Lower case	306	-.71	1.17
R	Upper case	433	-.65	.96
N	Upper case	1036	-.60	.41
p	Lower case	1036	-.60	.65
C	Upper case	297	-.54	1.45
m	Lower case	433	-.52	.31
D	Upper case	306	-.49	.67
P	Upper case	297	-.40	.82
n	Lower case	306	-.39	.40
F	Upper case	433	-.39	.32
f	Lower case	306	-.34	.33
I	Upper case	433	-.27	.85
K	Upper case	306	-.20	.54
k	Lower case	297	.05	.98
M	Upper case	297	.22	1.10
i	Lower case	306	.24	.77
c	Lower case	433	.26	.73
G	Upper case	1036	.31	.73
v	Lower case	297	.51	.96
z	Lower case	433	.65	.60
W	Upper case	306	.69	1.33
U	Upper case	297	.81	.80
h	Lower case	1036	.85	.77
Q	Upper case	297	.86	.98
u	Lower case	306	.88	.69
w	Lower case	433	.92	1.54
y	Lower case	730	.98	.62
l	Lower case	433	1.06	1.80

Table 2 (Continued)
Results of IRT Analysis of Letter Names Test, Spring Pilot, 2006

Letter	Upper or Lower Case	Count	Measure	Mean Square Outfit
V	Upper case	433	1.26	1.44
d	Lower case	306	1.85	1.36
J	Upper case	306	1.99	1.02
b	Lower case	297	2.14	1.79
j	Lower case	306	2.66	1.62
q	Lower case	433	7.02	1.46
g	Lower case	1036	unable to be estimated reliably	
H	Upper case	297	unable to be estimated reliably	
Y	Upper case	433	unable to be estimated reliably	

In all, 16 letters were outside the preferred Mean Square Outfit range. Seven (b, E, j, l, X, s, and w) exceeded a Mean Square Outfit of 1.50 while nine (A, a, F, f, N, n, m, O, and T) were below the recommended low of 0.50. In addition, the measure of 3 letters (g, H, and Y) was unable to be estimated (inability to estimate the measure occurs when individual items deviate from the pattern found in the rest of the test items to such an extent that the computer program is unable to calculate a reliable estimate of their difficulty).

Upon further examination of item information, we decided to include all 16 items whose Mean Square Outfit lay outside our ideal range of 0.50 – 1.50. The standard errors associated the estimate of the items' measure as well as the calculated measures for these items suggested that the items' estimate was sufficiently well-fit to allow us to use this information in constructing alternate forms of the Letter Names measure. The three more troublesome letters (g, H, and Y) for which no measure was able to be estimated, were excluded from our item bank and thus do not appear on any of our 20 alternate forms of the Letter Names measure.

Letter Sounds

The Letter Sounds items were piloted in the spring of 2006 using 5 common items across 3 separate forms of the measure to equate items across forms. Table 3 presents the results of this pilot testing.

Table 3
Results of IRT Analysis of Letter Sounds Test, Spring Pilot, 2006

Letter Sound	Upper or Lower Case	Count	Measure	Mean Square Outfit
D	Upper case	554	-3.32	3.59
m	Lower case	595	-2.77	.93
th	Lower case	554	-2.19	.65
Sh	Upper case	554	-2.00	.26
b	Lower case	1801	-1.84	.98
o	Lower case	554	-1.71	.48
k	Lower case	554	-1.67	.40
Ph	Upper case	652	-1.41	.48
c	Lower case	595	-1.38	.91
h	Lower case	1801	-1.23	1.22
e	Lower case	554	-1.17	1.19
Z	Upper case	595	-1.09	.32
Ch	Upper case	595	-1.06	1.47
U	Upper case	595	-1.06	.55
qu	Lower case	1801	-1.03	1.45
n	Lower case	595	-.92	.66
S	Upper case	652	-.88	2.53
T	Upper case	595	-.78	.44
f	Lower case	595	-.76	.70
I	Upper case	595	-.76	.24
M	Upper case	595	-.71	.37
H	Upper case	652	-.67	1.31
x	Lower case	554	-.60	1.86
z	Lower case	595	-.56	1.09
O	Upper case	652	-.56	1.69
sh	Lower case	554	-.56	1.76
wh	Lower case	595	-.50	.37
J	Upper case	595	-.41	.52
t	Lower case	595	-.39	.52
G	Upper case	595	-.37	1.48
N	Upper case	554	-.24	.70
l	Lower case	554	.00	1.30

Table 2 (Continued)
Results of IRT Analysis of Letter Sounds Test, Spring Pilot, 2006

Letter	Upper or Lower Case	Count	Measure	Mean Square Outfit
A	Upper case	652	.04	1.28
r	Lower case	1801	.33	1.11
L	Upper case	595	.36	1.23
y	Lower case	595	.57	.83
w	Lower case	652	.65	.93
v	Lower case	595	.91	1.12
Th	Upper case	595	.92	1.34
ch	Lower case	595	1.00	1.17
V	Upper case	554	1.02	1.19
a	Lower case	554	1.43	.68
E	Upper case	595	1.46	.96
g	Lower case	554	1.49	1.08
F	Upper case	595	1.49	1.33
ph	Lower case	652	1.54	.67
s	Lower case	595	1.55	1.58
i	Lower case	554	1.57	1.01
X	Upper case	554	2.05	.82
R	Upper case	595	2.13	1.04
Y	Upper case	554	2.65	1.27
K	Upper case	1801	2.66	1.03
u	Lower case	595	3.81	.79
P	Upper case	652	4.97	3.42

In all, 16 letters were outside the preferred Mean Square Outfit range. Seven (D, O, P, S, s, sh, and x) exceeded a Mean Square Outfit of 1.50 while nine (I, k, M, o, Ph , Sh, T, wh, and Z) were below the recommended low of 0.50. In addition, the measure of 6 letter sounds (B, C, d, j, p, and Qu) was unable to be estimated (inability to estimate the measure occurs when individual items deviate from the pattern found in the rest of the test items to such an extent that the computer program is unable to calculate a reliable estimate of their difficulty).

Upon further examination of item information, we decided to include all 16 items whose Mean Square Outfit lay outside our ideal range of 0.50 – 1.50. The standard errors associated the estimate of the items' measure as well as the calculated measures for these items suggested that the items' estimate was sufficiently well-fit to allow us to use this information in constructing

alternate forms of the Letter Sounds measure. The six more troublesome letter sounds (B, C, d, j, p, and Qu) for which no measure was able to be estimated, were excluded from our item bank and thus do not appear on any of our 20 alternate forms of the Letter Sounds measure.

Phoneme Segmenting

The Phoneme Segmenting items were piloted in the spring of 2006 using 5 common items across 20 separate forms of the measure to equate items across forms. Table 4 presents the results of this pilot testing.

Table 4
Results of IRT Analysis of Phoneme Segmenting Test, Spring Pilot, 2006

Word	Count	Measure	Mean Square Outfit	Standard Error
made	266	0.28	0.20	0.29
bane	220	0.29	0.27	0.29
nose	241	-1.70	0.29	0.20
net	266	0.38	0.32	0.24
boats	243	-0.99	0.33	0.18
boat	243	-0.99	0.33	0.18
pay	241	-2.27	0.33	0.30
phase	110	-0.38	0.34	0.29
knead	266	-1.28	0.35	0.29
crown	220	0.38	0.37	0.24
latch	237	-1.02	0.40	0.25
lab	110	-1.53	0.41	0.29
drag	220	-1.26	0.42	0.29
tip	243	-1.46	0.43	0.20
bit	243	-1.46	0.43	0.20
bat	237	-1.06	0.44	0.21
sheep	237	-1.18	0.46	0.20
rain	243	-1.02	0.48	0.17
vain	243	-1.02	0.48	0.17
chap	110	-1.27	0.50	0.28
bake	237	-1.21	0.53	0.20
paid	241	-2.02	0.53	0.20
lice	266	0.47	0.56	0.21
left	243	-0.74	0.57	0.17
chef	243	-0.74	0.57	0.17
bide	273	-0.14	0.58	0.17
cows	266	0.36	0.59	0.25
snout	266	1.18	0.60	0.12

Table 4 (Continued)
Results of IRT Analysis of Phoneme Segmenting Test, Spring Pilot, 2006

Word	Count	Measure	Mean Square Outfit	Standard Error
scan	266	0.09	0.61	0.13
loaf	266	0.46	0.63	0.22
float	237	0.43	0.64	0.12
nab	266	0.33	0.65	0.26
home	237	-0.35	0.65	0.16
snake	243	0.40	0.66	0.12
hire	243	-0.92	0.66	0.18
clown	243	0.40	0.66	0.12
dive	243	-0.92	0.66	0.18
read	110	-1.14	0.66	0.28
roach	220	0.48	0.67	0.21
soak	273	-0.46	0.67	0.19
green	241	-0.12	0.68	0.13
glum	243	0.32	0.69	0.12
sled	243	0.32	0.69	0.12
glows	243	-0.92	0.70	0.17
down	243	-0.92	0.70	0.17
gift	220	0.37	0.71	0.25
moat	234	1.69	0.71	0.16
sneak	241	-0.29	0.71	0.13
lump	110	0.40	0.71	0.19
nurse	243	-1.78	0.72	0.21
slap	237	0.40	0.72	0.12
mom	243	-1.78	0.72	0.21
pounce	241	-0.43	0.72	0.13
rant	266	0.09	0.73	0.13
glitch	266	1.66	0.73	0.11
wren	243	0.37	0.73	0.13
knives	243	0.37	0.73	0.13
slab	241	-0.40	0.73	0.13
snare	220	0.10	0.74	0.13
snail	273	-0.19	0.74	0.13
snag	266	1.17	0.75	0.13
graph	266	0.32	0.75	0.12
lend	220	0.47	0.75	0.22
black	237	0.92	0.75	0.13
drip	237	0.52	0.76	0.12
wrist	241	-0.78	0.76	0.13
crew	110	-0.30	0.76	0.21
main	266	-1.21	0.77	0.27

Table 4 (Continued)
Results of IRT Analysis of Phoneme Segmenting Test, Spring Pilot, 2006

Word	Count	Measure	Mean Square Outfit	Standard Error
rank	243	0.32	0.77	0.13
fit	234	-1.42	0.77	0.18
dunk	243	0.32	0.77	0.13
boast	110	-0.16	0.77	0.19
lamb	273	-0.39	0.78	0.18
trait	110	0.52	0.78	0.17
span	243	0.48	0.79	0.12
blood	234	-0.05	0.79	0.11
wind	237	0.21	0.79	0.13
spill	243	0.48	0.79	0.12
lag	220	0.33	0.80	0.26
desk	241	-0.51	0.80	0.13
bent	241	-0.6	0.80	0.14
trip	273	0.00	0.81	0.12
jobless	243	0.05	0.81	0.13
metal	243	0.05	0.81	0.13
owner	110	0.39	0.81	0.22
bold	266	1.75	0.82	0.11
clink	110	1.14	0.82	0.15
nine	2067	-1.17	0.83	0.07
smile	273	0.15	0.83	0.12
mess	273	-0.12	0.83	0.16
shout	243	-0.70	0.83	0.17
shirt	234	-1.28	0.83	0.17
foul	243	-0.70	0.83	0.17
skin	241	-0.08	0.83	0.12
kettle	266	1.04	0.84	0.14
first	273	0.41	0.84	0.11
brace	243	1.24	0.84	0.10
omen	243	1.15	0.84	0.11
must	234	-0.29	0.84	0.12
brand	243	1.24	0.84	0.10
open	243	1.15	0.84	0.11
tap	2066	-1.89	0.85	0.08
globe	266	1.26	0.86	0.13
lime	220	0.10	0.86	0.13
roman	273	0.76	0.86	0.09
cane	234	-0.22	0.87	0.16
inspire	220	1.66	0.88	0.11
lame	273	-0.22	0.88	0.18
shade	273	-0.88	0.88	0.22
shiny	273	0.11	0.88	0.13

Table 4 (Continued)
Results of IRT Analysis of Phoneme Segmenting Test, Spring Pilot, 2006

Word	Count	Measure	Mean Square Outfit	Standard Error
jump	234	0.53	0.88	0.12
flowing	237	1.16	0.88	0.10
yam	266	0.53	0.89	0.20
trap	273	0.46	0.89	0.11
crowd	273	0.76	0.89	0.11
wear	273	0.07	0.90	0.14
scale	237	0.37	0.90	0.12
glass	220	1.18	0.91	0.13
spoken	220	-1.20	0.91	0.27
crumb	234	0.29	0.91	0.11
dimmer	241	-0.30	0.91	0.13
sealer	241	-0.04	0.91	0.12
stack	273	-0.06	0.92	0.11
found	237	0.36	0.92	0.13
hound	110	-0.06	0.94	0.18
silent	266	1.34	0.95	0.09
hour	220	0.33	0.95	0.12
spouse	273	0.51	0.95	0.11
fold	237	0.46	0.95	0.12
ramp	220	1.75	0.96	0.11
draw	266	-0.53	0.97	0.16
bend	234	0.28	0.97	0.11
treated	233	1.48	0.97	0.08
letter	237	0.46	0.97	0.12
jar	241	-0.91	0.97	0.16
lust	110	-0.40	0.97	0.19
straight	273	1.30	0.98	0.08
cleanest	273	1.75	0.98	0.07
rule	243	-1.27	0.98	0.19
scoop	234	0.20	0.98	0.11
lobe	243	-1.27	0.98	0.19
remote	241	0.73	0.98	0.10
regrow	110	0.86	0.98	0.15
release	266	1.88	0.99	0.10
cup	2067	-1.90	1.00	0.08
sneaky	220	1.05	1.00	0.14
mint	234	-0.24	1.00	0.12
tint	237	0.56	1.01	0.13
wraps	237	0.03	1.01	0.13
bunk	220	1.26	1.02	0.13
Bone	233	-0.21	1.02	0.16
jokes	273	0.09	1.03	0.12

Table 4 (Continued)
Results of IRT Analysis of Phoneme Segmenting Test, Spring Pilot, 2006

Word	Count	Measure	Mean Square Outfit	Standard Error
sped	237	0.41	1.03	0.12
pack	241	-1.70	1.03	0.19
race	110	-0.95	1.03	0.29
fray	110	1.14	1.03	0.22
theft	266	1.35	1.04	0.13
word	220	0.54	1.04	0.20
gin	237	-0.22	1.05	0.16
send	110	0.42	1.05	0.18
mass	234	-1.02	1.06	0.16
blur	241	-0.02	1.06	0.14
box	234	-0.40	1.07	0.13
slowly	234	0.90	1.07	0.09
include	243	1.76	1.10	0.10
bow	234	2.12	1.10	0.25
apron	243	1.76	1.10	0.10
male	241	-1.59	1.10	0.17
mean	237	2.74	1.11	0.17
mine	241	-1.59	1.11	0.19
repeal	110	0.87	1.11	0.14
leaping	237	1.25	1.12	0.11
gnat	273	-0.91	1.13	0.20
listen	234	1.04	1.13	0.09
rental	220	1.34	1.14	0.09
rack	273	-0.23	1.14	0.18
free	220	-0.52	1.16	0.16
shed	241	-1.59	1.16	0.19
street	234	0.92	1.17	0.09
strap	110	1.20	1.17	0.14
city	2067	-0.19	1.18	0.04
raid	220	1.88	1.18	0.10
then	220	1.35	1.21	0.12
seal	273	-0.28	1.24	0.16
neater	243	1.29	1.24	0.10
repeat	243	1.29	1.24	0.10
thoughtless	273	1.38	1.26	0.07
able	234	0.26	1.26	0.12
huddle	110	0.46	1.27	0.17
omit	266	1.55	1.29	0.11
ouch	234	-0.78	1.31	0.20
hid	241	-1.89	1.31	0.19
futile	241	1.66	1.34	0.10
these	241	-0.50	1.37	0.16

Table 4 (Continued)
Results of IRT Analysis of Phoneme Segmenting Test, Spring Pilot, 2006

Word	Count	Measure	Mean Square Outfit	Standard Error
fair	266	-0.55	1.38	0.22
chalk	243	-0.51	1.39	0.16
theme	243	-0.51	1.39	0.16
tail	234	-0.79	1.40	0.15
rear	234	0.11	1.44	0.13
wrath	234	-0.50	1.50	0.15
knots	220	1.55	1.54	0.11
rude	110	-1.25	1.54	0.27
tin	110	-1.07	1.60	0.25
love	220	-0.54	1.63	0.22
ode	266	0.85	1.70	0.23
oath	241	-0.91	1.70	0.22
maid	237	-0.68	1.71	0.18
brat	243	-1.61	1.77	0.20
sad	243	-1.61	1.77	0.20
nut	273	-1.01	1.82	0.27
kite	243	-0.91	1.84	0.18
yard	243	-0.91	1.84	0.18
wing	243	0.59	1.97	0.13
aunt	243	0.59	1.97	0.13
game	243	-0.68	2.05	0.21
gate	243	-0.68	2.05	0.21
ripe	237	2.61	2.11	0.18
bean	220	0.86	2.13	0.23
five	110	-1.01	2.19	0.26
arrow	237	0.77	2.53	0.13
show	1847	4.27	2.64	0.11
seam	243	-0.63	2.99	0.21
ease	243	-0.63	2.99	0.21
sow	110	-0.74	3.28	0.45
rise	234	-0.96	4.10	0.16
fumes	220	-1.90	4.15	0.93
jaw	110	-2.23	5.06	0.44
joy	237	-1.11	5.56	0.26
toy	241	-1.77	8.56	0.26

In all, 48 words were outside the preferred Mean Square Outfit range, with 29 exceeding a Mean Square Outfit of 1.50 and 19 below the recommended low of 0.50.Exclusion of these words from our final item bank resulted in a total of 181 words remaining in the item bank.

Discussion

We begin with a discussion each of the measures. One alternate forms of the measures were created, all forms of the measures were then loaded to the EasyCBM website (easycbm.com) for web-based access.

Using the Results of the Pilot Testing to Create Alternate Forms of the Letter Names Measures

Using results of the pilot testing, we clustered all Letter Names that were able to be estimated into three categories: easy, moderate and difficult (see Table 5). We used this information to draw items to create 20 alternate forms of the Progress Monitoring measures. In all cases, we drew from the easy items for the first two rows of items, the moderate items for the two middle rows, and the difficult items for the final two rows of items.

Table 5
Clustering Letter Names into Three Categories of Difficulty

Easy Items	Moderate Items	Difficult Items
o	R	v
X	N	z
A	p	W
s	C	U
O	m	h
B	D	Q
E	P	u
a	n	w
T	F	y
x	f	l
e	I	V
r	K	d
Z	k	J
S	M	b
L	i	j
t	c	q
	G	

We arranged these items on 20 alternate forms of the Letter Names measure, assigning the first 20 forms to Kindergarten, then going through and assigning the next 20 forms to first grade. This process resulted in 20 comparable forms at each of those two grade levels. For the

Student Form of the measures, we used size 28 Comic Sans MC font (see Appendix A). The Assessor Copy of each of the forms includes administration and scoring directions as well as a smaller version of the student measure (see Appendix B). All forms of the measures were then loaded to the EasyCBM website for web-based access.

Using the Results of the Pilot Testing to Create Alternate Forms of the Letter Sounds Measures

Using results of the pilot testing, we clustered all Letter Sounds that were able to be estimated into three categories: easy, moderate and difficult (see Table 6). We used this information to draw items to create 20 alternate forms of the Progress Monitoring measures. In all cases, we drew from the easy items for the first two rows of items, the moderate items for the two middle rows, and the difficult items for the final two rows of items.

Table 6
Clustering Letter Sounds into Three Categories of Difficulty

Easy Items	Moderate Items	Difficult Items
D	f	w
m	I	v
th	M	Th
Sh	H	ch
b	x	V
o	z	a
k	O	E
Ph	sh	g
c	wh	F
h	J	ph
e	t	s
Z	G	i
Ch	N	X
U	l	R
qu	A	Y
n	r	K
S	L	u
T	y	P

We arranged these items on 20 alternate forms of the Letter Sounds measure, assigning the first 20 forms to Kindergarten, then going through and assigning the next 20 forms to first

grade. This process resulted in 20 comparable forms at each of those two grade levels. For the Student Form of the measures, we used size 28 Comic Sans MC font (see Appendix C). The Assessor Copy of each of the forms includes administration and scoring directions as well as a smaller version of the student measure (see Appendix D). All forms of the measures were then loaded to the EasyCBM website for web-based access.

Using Pilot Results to Create Alternate Forms of the Phoneme Segmenting Measures

Using results of the pilot testing, we clustered all remaining items into 14 different categories representing different levels of item difficulty (see Table 7). We used this information to draw items used to create 20 alternate forms of the Phoneme Segmenting measures for use in Kindergarten and 20 alternate forms for use in first grade.

Table 7

Phoneme Segmenting Item Bank, Clustered into 14 Categories of Difficulty

Category	Average Measure	Items		
1	-1.74	paid	tap	pack
		cup	mom	male
		hid	nurse	mine
		shed	fit	
2	-1.08	shirt	nine	hire
		chap	read (ee)	gnat
		lobe	mass	jar
		rule	race	shade
		bake	dive	glows
		main	down	spoken

Table 7 (Continued

Phoneme Segmenting Item Bank, Clustered into 14 Categories of Difficulty

Category	Average Measure	Items		
3	-0.62	tail	left	fair
		ouch	foul	draw
		wrist	shout	free
		chef	bent	chalk
		desk	theme	these
		wrath		
4	-0.36	soak	slab	dimmer
		pounce	lamb	must
		box	home	sneak
		lust	crew	seal
5	-0.17	mint	lame	boast
		rack	bone	bide
		cane	city	green
		gin	snail	mess
		skin	hound	
6	0.05	stack	jobless	lime
		blood	metal	snare
		sealer	wear	rear
		blur	jokes	shiny
		trip	rant	scan
		wraps		

Table 7 (Continued

Phoneme Segmenting Item Bank, Clustered into 14 Categories of Difficulty

Category	Average Measure	Items		
7	0.28	smile	crumb	sled
		scoop	dunk	hour
		wind	glum	lag
		able	graph	nab
		bend	rank	
8	0.42	cows	wren	snake
		found	owner	first
		gift	clown	sped
		knives	lump	send
		scale	slap	float
		fold	loaf	lice
		huddle	trap	roach
		letter	lend	span
		spill		
9	0.70	spouse	word	regrow
		drip	tint	repeal
		trait	remote	slowly
		jump	crowd	black
		yam	roman	street

Table 7 (Continued

Phoneme Segmenting Item Bank, Clustered into 14 Categories of Difficulty

Category	Average Measure	Items		
10	1.13	kettle	fray	snag
		listen	omen	glass
		sneaky	open	snout
		clink	flowing	
11	1.29	strap	globe	silent
		brace	neater	theft
		brand	repeat	then
		leaping	straight	thoughtless
		bunk	rental	
12	1.62	treated	futile	inspire
		omit	glitch	moat
13	1.79	bold	apron	raid
		cleanest	include	release
		ramp		
14	2.43	bow	mean	

Because the Phoneme Segmenting measures are administered entirely orally, no student versions of these forms was created. The Assessor Copy of each of the forms includes administration and scoring directions as well as the items test administrators use with the students (see Appendix E).

Appendix A

Example Letter Names Test: Student Copy

Letter Names

o	X	A	s	O	B	E	a	T	x
e	r	Z	S	L	t	R	N	p	C
m	D	P	n	F	I	M	f	K	i
k	c	G	v	z	W	U	h	Q	u
w	y	l	V	d	J	b	j	q	A

Appendix B

Example Letter Names Test: Assessor Copy

Student Name: _____ Student ID #: _____

Teacher Name: _____ School: _____

Letter Names

Procedures

Place the probe marked "Letter Names Student Copy" in front of the student. Read the directions to the student. When you are finished administering the test, enter the student results on the website for scoring and record keeping.

Directions

"When I say begin, say the name of each letter. I will stop you after 30 seconds. Start at the top of the page and read across each row." Demonstrate by sweeping your finger from left to right across the first row. **"Move your marker down after each row."** Demonstrate. **"Any questions?... Ready?...Begin."** At 30 seconds, say "**Stop."** Mark the last letter with a bracket.]

Note: This is a 30 second timed test.

Scoring

If student:

- Self corrects, write S.C. above letter name and count as correct.
- Says incorrect letter name, slash through letter name, and count as incorrect.
- Hesitates more than 3 seconds, supply the letter name and count as incorrect.
- Skips letter, circle the letter and count as incorrect.
- Clearly loses his/her place, point to the next letter.

o	X	A	s	O	B	E	a	T	x	10
e	r	Z	S	L	t	R	N	p	C	20
m	D	P	n	F	I	M	f	K	i	30
k	c	G	v	z	W	U	h	Q	u	40
w	y	l	V	d	J	b	j	q	A	50

 # Correct _____

Appendix C

Example Letter Sounds Test: Student Copy

Letter Sounds

D	m	th	Sh	b	o	k	Ph	c
h	e	Z	Ch	U	qu	n	S	T
f	I	M	H	x	z	O	sh	wh
J	t	G	N	l	A	r	L	y
w	v	Th	ch	V	a	E	g	F
ph	s	i	X	R	Y	K	u	P

Appendix D

Example Letter Sounds Test: Assessor Copy

Student Name: _____ Student ID #: _____

Teacher Name: _____ School: _____

Letter Sounds

Procedures

Place the probe marked "Letter Sounds Student Copy" in front of the student. Read the directions to the student. When you are finished administering the test, enter the student results on the website for scoring and record keeping.

Directions

"When I say begin, say the sound each letter makes. I will stop you after 30 seconds. Start at the top of the page and read across each row." Demonstrate by sweeping your finger from left to right across the first row. **"Move your marker down after each row."** Demonstrate. **"Any questions?....Ready?...Begin."** At 30 seconds, say "**Stop."** Mark the last letter with a bracket.]

Note: This is a 30 second timed test.

Scoring

If student:

- Says letter name instead of sound, say **"Can you tell me what sound that letter makes?"** If student says letter name again, count as incorrect.
- Self corrects, write S.C. above letter sound and count as correct.
- Says incorrect letter sound, slash through letter and count as incorrect.
- Hesitates more than 3 seconds, supply the letter sound and count as incorrect.
- Skips letter, circle the letter and count as incorrect.
- Clearly loses his/her place, point to the next letter.

D	m	th	Sh	b	o	k	Ph	c	9
h	e	Z	Ch	U	qu	n	S	T	18
f	I	M	H	x	z	O	sh	wh	27
J	t	G	N	l	A	r	L	y	36
w	v	Th	ch	V	a	E	g	F	45
ph	s	i	X	R	Y	K	u	P	54

 # Correct _____

Appendix E

Example Phoneme Segmenting Test: Assessor Copy

Phoneme Segmentation (1_2)

Student: _____ School: _____

Grade: _____ Date: _____ Assessor: _____

This test is administered entirely orally. Do NOT show the student this scoring sheet.

Say to student: "**I am going to say a word, and you will give me the sounds you hear in that word. If I say *cap*, you will say /c/ /a/ /p/. If I say *it*, you will say /i/ /t/. If I say *top*, you will say /t/ /o/ /p/. Let's try it."**

Give the student 3 practice trials using *no*, *club*, and *ten*. After each response, provide student feedback by saying '**correct**' or '**incorrect**.' For incorrect responses, give student the correct response before going to the next practice item. After the three trials, begin the test.

Note: This test is timed for 60 seconds.

Item	Teacher Says	Student Says	Number Correct	Item	Teacher Says	Student Says	Number Correct
1	pack	/p/ /a/ /ck/	___ / 3	**16**	sneaky	/s/ /n/ /ea/ /k/ /y/	___ / 5
2	main	/m/ /ai/ /n/	___ / 3	**17**	silent	/s/ /i/ /l/ /e/ /n/ /t/	___ / 6
3	desk	/d/ /e/ /s/ /k/	___ / 4	**18**	omit	/o/ /m/ /i/ /t/	___ / 4
4	lamb	/l/ /a/ /mb/	___ / 3	**19**	release	/r/ /e/ /l/ /ea/ /se/	___ / 5
5	bone	/b/ /o/ /ne/	___ / 3	**20**	bow (ou)	/b/ /ow/	___ / 2
6	skin	/s/ /k/ /i/ /n/	___ / 4	**21**	paid	/p/ /ai/ /d/	___ / 3
7	wraps	/wr/ /a/ /p/ /s/	___ / 4	**22**	shirt	/sh/ /ir/ /t/	___ / 3
8	graph	/g/ /r/ /a/ /ph/	___ / 4	**23**	dive	/d/ /i/ /ve/	___ / 3
9	wren	/wr/ /e/ /n/	___ / 3	**24**	left	/l/ /e/ /f/ /t/	___ / 4
10	black	/b/ /l/ /a/ /ck/	___ / 4	**25**			

Total Number Correct ___ / 70

References

Alonzo, J., & Tindal, G. (2007). *Examining the Technical Adequacy of Word and Passage Reading Fluency Measures in a Progress Monitoring Assessment System (Technical Report # 40).* Eugene, OR: Behavioral Research and Teaching.

Alonzo, J., Liu, K., & Tindal, G. (2007). *Examining the Technical Adequacy of Reading Comprehension Measures in a Progress Monitoring Assessment System (Technical Report # 41).* Eugene, OR: Behavioral Research and Teaching.

Alonzo, J., Tindal, G., & Ketterlin-Geller, L.R. (2007). *General outcome measures of basic skills in reading and math*. In L. Florian (Ed.), Handbook of Special Education. Thousand Oaks, CA: Sage.

Deno, S. L., & Mirkin, P. M. (1977). *Data based program modification.* Minneapolis, MN: University of Minnesota Leadership Training Institute/Special Education.

Ehri, L.C. (1991). Development of the ability to read words. In R. Barr, M.L. Kamil, P.B. Mosenthal, & P.D. Pearson (Eds.) *Handbook of reading research, Volume II.* New York: Longman.

Ehri, L.C. (2005). Learning to read words: Theory, findings, and issues. *Scientific Studies of Reading, 9,* 167-188.

Graves, A. W., Plasencia-Peinando, J., Deno, S. L., & Johnson, J. R. (2005). Formatively evaluating the reading progress of first-grade English learners in multiple-language classrooms. *Remedial & Special Education, 26,* 215-225.

Hasbrouck, J. & Tindal, G. (2005). Oral reading fluency norms: a valuable tool for reading teachers. *The Reading Teacher*.

Kolen, M. J. & Brennan, R. L., (1995). *Test equating: Methods and practices*. New York: Springer.

Linacre, J. M. (2006). Winsteps Rasch Measurement, version 3.61.1. Author.

National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction.* Washington, DC: Author.

Reading First. (2006). U.S. Department of Education. Retrieved August 8, 2006 from http://www.ed.gov/programs/readingfirst/index.html

Ritchey, K. D., & Speece, D. L. (2006). From letter names to word reading: The nascent role of sublexical fluency. *Contemporary Educational Psychology, 31*, 301-327.